Modeling two-dimensional infrared spectroscopy of hydrogen bonded systems

De Carvalho Vicente Da Cunha, Ana

4 ASSESSING SPECTRAL SIMULATION PROTOCOLS FOR THE AMIDE I BAND OF PROTEINS[1]

We present a benchmark study of spectral simulation protocols for the amide I band of proteins. The amide I band is widely used in infrared spectroscopy of proteins due to the large signal intensity, high sensitivity to hydrogen bonding, and secondary structural motifs. This band has, thus, proven valuable in many studies of protein structure function relationships. We benchmark spectral simulation protocols using two common force fields in combination with several electrostatic mappings and coupling models. The results are validated against experimental linear absorption and two-dimensional infrared spectroscopy (2D IR) for three well-studied proteins. We find 2D IR to be much more sensitive to the simulation protocol than linear absorption and report on the best simulation protocols. The findings demonstrate that there is still room for ideas to improve the existing models for the amide I band of proteins.

[1] THIS CHAPTER WAS PUBLISHED IN J. CHEM. THEORY COMPUT.,12 (8),3982-3992,2016.

4.1 INTRODUCTION

Conformational changes in proteins are related to protein function. Such fluctuations happen on time scales ranging from the picosecond to the millisecond timescale and beyond. Fast fluctuations have been shown to influence long timescale processes [12–20]. Probing fast dynamics along the reaction coordinate can, thus, provide deeper knowledge of structure-activity relationships of biological systems. [67–75]. The amide I band is the most probed mode in IR spectroscopy of proteins, due to its sensitivity to solvation and secondary structure. It is dominated by the CO stretch vibrations in the peptide backbone, and located in the frequency range from 1600 to 1700 cm^{-1}. The most common signatures of secondary structure in the amide I band are peaks between 1630 and 1640 cm^{-1}, and between 1640 and 1650 cm^{-1}, resulting from β-sheets and α-helices, respectively, Figure 4.1. 2D IR [23] is a novel technique to probe transient structure and dynamics.

It has been applied extensively to the amide I mode of protein systems [23,30,37,72,78–80,91–107], providing structural and dynamical information that conventional absorption spectroscopy is not sensitive to. The main limitation of infrared spectroscopy is the lack of distinct peaks in such spectra arising from a broad distribution of different hydrogen bonding environments, delocalization of vibrational modes [79], and side chain absorption in the amide I region [29, 115]. Therefore, theoretical models [24–33] have been developed to disentangle spectral signals, and allow interpretation in terms of structure and dynamics of the investigated proteins. So far, few thorough tests of the theoretical models have been made. The existing benchmarks have generally been limited to peptides [28, 116–118] or proteins with isotope labeled sites [43, 81]. The aim of this work is to benchmark some of the most popular existing models using linear absorption and 2D IR spectroscopy on full proteins to guide the choice of simulation protocol in future studies and identify limitations of the existing methods.

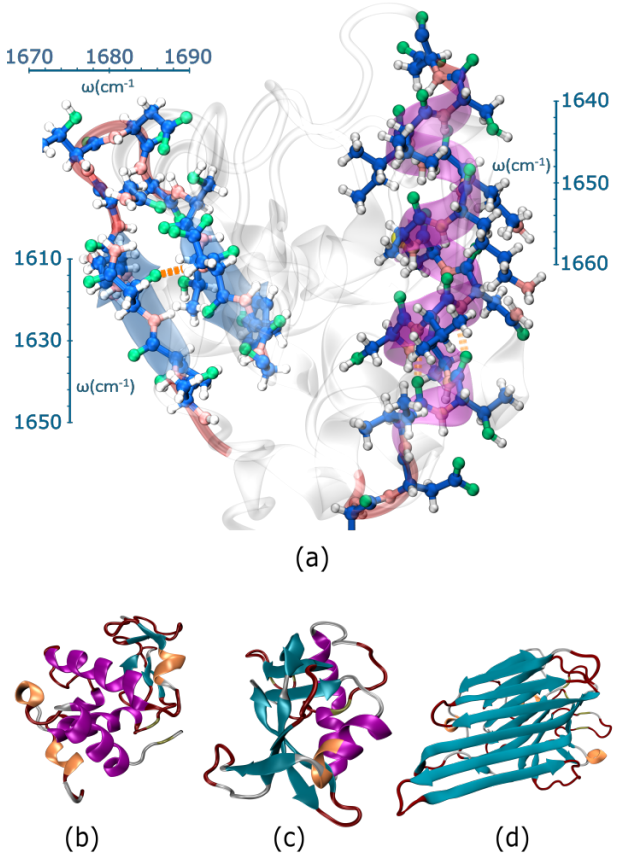

Figure 4.1: (a) The relationship between secondary structure and amide I vibrational frequency in the infrared spectrum. Where the residues of, both, α-helix and β-sheet are in atomistic representation, and the overall protein structure is represented in new cartoon. The oxygens are shown in green, the nitrogens in pink, the carbons in blue, and the hydrogens in light grey. An α-helix is colored in purple, and a β-sheet in blue. The hydrogen bonds between different residues are represented in orange. The structures of the used proteins are represented in new cartoon (b) Lysozyme, (c) Ribonuclease A, and (d) Concanavalin A, rendered with VMD-1.9.2 [119]. Here the random coils are presented in red, and the π-helix in orange.

The main advantage of 2D IR spectroscopy is that the vibrational spectrum is spread over two frequency axes relating the frequency of an initial excitation of a given vibration with the remaining vibrations and, thus, sensitive to vibrational couplings [23, 91]. It is, thus, well documented that this spectroscopy is much more sensitive to structural changes than conventional linear spectroscopy. The 2D IR experiment is performed by sending a number of femtosecond laser pulses through the sample and detecting the third-order response. The first laser pulse brings the system into a coherent superposition of the ground state and the single excited state. The second laser pulse can create a coherent superposition of two single excited states, or can de-excite the system back to the ground state, or it can create a single excited state population. The third laser pulse either brings the system to a coherent superposition between the ground state and a single excited state, or between double excited state. Finally, a time domain signal is emitted, and posteriorly Fourier transformed with respect to the time delay between the two first pulses, t_1, and the time between the last pulse and the signal detection, t_3, to the frequency domain. The time delay between the second and the third pulses is denoted the waiting time and may be used to study dynamics. In the present work the waiting time will be zero. The interactions of the system with the laser applied pulses results in three processes: (a) ground state bleach, (b) stimulated emission, and (c) excited state absorption, and are represented by the Feynmann diagrams shown in Figure 3.8 [89].

The 2D IR spectra are given by the sum of the signal emitted in two directions $\vec{k}^I = -\vec{k}_1 + \vec{k}_2 + \vec{k}_3$ and $\vec{k}^{II} = \vec{k}_1 - \vec{k}_2 + \vec{k}_3$. Here, $\vec{k}_1$, $\vec{k}_2$, and $\vec{k}_3$ are the wave vectors on the incoming fields, and $\vec{k}^I$, is the photon echo, whereas $\vec{k}^{II}$ is the non-rephasing signal. An echo signal is produced in the case of the rephasing diagram, due to the fact that the phase coherent oscillation during t_3 and t_1 are opposite. The 2D IR spectrum, thus, contain several peaks, where the diagonal ones correspond to the ground state bleach and stimulated emission, which both result in decreased absorption. Below the diagonal excited state absorption peaks are present, the process results in an additional absorption, and, these peaks have the opposite sign of the diagonal peaks. Furthermore, cross peaks, arise due to the coupling between different modes. This occurs due to the delocalization of vibrational excitons [91, 105, 109].

The protocols for simulating the amide I spectra of proteins considered in this Chapter follow the same basic procedure outlined below. First classical MD simulations [84–86] are performed to determine the structure and dynamics of the protein under investigation. The time-dependent Hamiltonian accounting for

the amide I vibrations is constructed for numerous snapshots along the trajectory. This is achieved by using mappings that relate the electrostatic environment predicted by the force field with the local mode vibrational frequencies [27, 27–29, 32, 34–43], and couplings [38, 120–123]. This information is then converted to linear absorption and 2D IR spectra using response function based calculations [82, 83]. The force field point charges determining the electric fields were not parametrized for this type of modelling, but some of the mappings were developed for specific force fields [29, 37]. For not matching combinations one can, thus, expect significant errors for other force fields than that they we parameterized for. Furthermore, within the wide range of existing force fields, some do not contain parameters for various coenzymes, narrowing the options of usage or requiring new parameterizations.

Benchmarking studies of the amide I, so far, focused on linear absorption spectra in solution [32,124], single conformation gas phase spectroscopy [117,118], and 2D IR spectroscopy utilizing isotope labels [28,43,81]. In a recent paper [125], a benchmark study using linear absorption and 2D IR for full proteins was performed. That study focused on the Gromos-54a7 and Amber99SB-ILDN force fields. Three well studied proteins, namely Lysozyme [126] (Lys, Protein Data Bank Identication (PDB ID): 1AKI), Ribonuclease A [127] (RNseA, PDB ID: 1FS3), and Concanavalin A [128] (ConA, PDB ID: 1NLS) shown in Figure 4.1 were used. These were chosen due to their varying α-helix and β-sheet content. This strategy allowed the determination of quantitative measures for the performance of different simulation protocols. We will follow the same benchmarking strategy of that paper, but focus on the CHARMM-27 [129] and OPLS-AA [130] force fields. These are generally popular protein force fields, which have both previous been used for modelling the amide-I band. We will therefore be able to make a direct comparison between results for all four force fields. As the present benchmarking procedure is computationally time very demanding we limit our study to a limited set of simulation protocols using the knowledge of previous benchmarking studies [28, 43, 81, 116–118] to narrow down the selected protocols.

In the current Chapter, a benchmark of different electrostatic mappings, and coupling models, are presented. The main goal is to provide information of the best combination of force fields, electrostatic mappings, and coupling models. The detailed simulation and benchmarking protocol are presented in the Methods section. In the Results section the obtained FTIR and 2D IR spectra and presented and quantitatively compared with experiment. The conclusions are drawn in section IV.

4.2 METHODS

4.2.1 Molecular dynamics

All molecular dynamics simulations were performed with the GROMACS [131] suite 4.6.1 using the OPLS-AA [130] and the CHARMM-27 [129] force fields. All protein systems were solvated with water using the SPC/E model [132], and sodium/chlorine counter ions were added to keep the simulation box neutral. At first an energy minimization was made, followed by two 100 ps equilibrations: (1) an NVT equilibration at a temperature of 300 K [133]; and (2) an NPT at pressure of 1 bar using a Parrinello-Rahman barostat [134]. A constant volume production run of 1 ns, at 300 K, with a time step of 2 fs was performed, in which a 1.0 nm cutoff was used for both Lennard-Jones and Coulomb interactions [135]. The production runs were performed in the NVT ensemble to avoid possible artifacts due to dynamical coupling between the systems and the barostat. The long range Coulomb interactions were treated using the Particle Mesh Ewald (PME) method [136], with a grid step of 0.16 nm, and a convergence of 10^{-5}. The truncation of Lennard-Jones interactions was compensated by introducing analytic corrections to the pressure and potential energy [135]. A V-rescale thermostat [133] with an inverse time constant of τ^{-1}= 0.2 ps^{-1} was used to keep the temperature constant. All bonds were constrained using the LINCS algorithm [137]. Trajectories of all atomic positions were stored with 20 fs intervals between the snapshots, for the spectral calculations.

4.2.2 The Amide I Hamiltonian

The time dependent Hamiltonian for the amide I modes was constructed from all stored snapshots of the MD production run:

$$H(t) = \sum_i \omega_i(t) B_i^\dagger B_i - \sum_i \frac{\Delta}{2}(t) B_i^\dagger B_i^\dagger B_i B_i \qquad (4.1)$$
$$+ \sum_{i,j} J_{ij}(t) B_j^\dagger B_i - \sum_i \vec{\mu_i}(t) \cdot \vec{E}(t)(B_i^\dagger + B_i)$$

Here $B_i^\dagger$ and B_i are the bosonic creation and annihilation operators, and $\vec{E}$(t) is the external laser field used to excite the amide I units. The site frequencies, ω_i(t), the transitions dipoles, $\vec{\mu_i}$(t), of each amide unit of the protein backbone were calculated with electrostatic maps. The anharmonicity Δ was kept constant

at 16 cm^{-1}. [23] The electrostatic maps relate the electric fields created by the force field point charges on each atom of the amide I unit, with the site frequency and the transition dipole. The maps all assume dependence of the frequencies and transition dipoles on the electric field, and in our case the electric field gradient as well. In this study three electrostatic maps were applied: the Skinner map [29], the Jansen map [36], and the Tokmakoff map [30]. The Skinner and the Tokmakoff maps were parametrized for Gromos-54a7 [138], and for CHARMM-27 [129], respectively. The map coefficients were fitted empirically with the charges generated by the force fields, either using *ab initio* results or experiment on the dipeptide for the fit. Here we will only use the Tokmakoff map for the CHARMM-27 force field for which it was developed. The Jansen frequency map [36] accounts for dependence on both, electric field generated by the surroundings, and its gradient. This map was constructed with DFT calculations in 75 different point charge environments, and is thus, not optimized for any particular force field. In all cases, the short range couplings and frequency shifts were calculated using the nearest neighbour coupling model. This is a Ramachandran angle based mapping parametrized from DFT calculations on dipeptides [120, 123, 139]. The long range couplings between amide units at larger distances, J_{ij}, were calculated using either the transition charge coupling (TCC) [120], or the transition dipole coupling model (TDC) [121].

For the TDC model, Eq. (4.2), the coupling is determined from the transition dipoles, $\vec{\mu_i}$, as taken from Ref. 121, of two involved units, i and j, where $\vec{r}_{ij}$ is distance between the transition dipoles [29].

$$J_{ij} = \frac{1}{4\pi\varepsilon_0}\left(\frac{\vec{\mu_i}.\vec{\mu_j}}{r_{ij}^3} - 3\frac{(\vec{\mu_i}.\vec{r_{ij}})(\vec{\mu_i}.\vec{r_{ij}})}{r_{ij}^5}\right) \tag{4.2}$$

The TCC coupling has the form [120, 140]:

$$J_{ij} = \frac{1}{4\pi\varepsilon_0}\sum_{n,m}\Big(\frac{dq_n dq_m}{|\vec{r}_{n_i m_j}|} - \frac{3q_n q_m(\vec{\nu}_{n_i}.\vec{r}_{n_i m_j})(\vec{\nu}_{m_j}.\vec{r}_{n_i m_j})}{|\vec{r}_{n_i m_j}|^5} - \frac{-dq_n q_m \vec{\nu}_{m_j}.\vec{r}_{n_i m_j} + q_n dq_m \vec{\nu}_{n_i}.\vec{r}_{n_i m_j} - q_n q_m \vec{\nu}_{n_i}.\vec{\nu}_{m_j}}{|\vec{r}_{n_i m_j}|^3}\Big) \tag{4.3}$$

Here, the subscripts n and m number the atoms belonging to different amide units i, and j, respectively. A charge, q_n, a transition charge, dq_n, and a normal mode coordinate, $\vec{\nu_i}$, are assigned to each atom of each amide I unit, of the

protein backbone. $\vec{r}_{n_i m_j}$ is the distance vector between atoms of the two units. The used parameters are taken from Refs. 120, 123 and from Ref. 139 for the units preceding proline.

4.2.3 Spectral Calculations

All spectra were calculated with the Numerical Integration of Schrödinger Equation method (NISE) [83, 108], in which the time dependent Schrödinger equation is solved numerically for the Amide I time dependent Hamiltonian. Here, an instantaneous interaction between an external field and the system is assumed, and when the electric field is vanishing the coupling between states with different excitation is zero. Therefore, the Hamiltonian is block diagonal with a ground state block (g), an excited state block (e), and double excited state block (f). Consequently these blocks can be treated separately [83, 108]. To solve the time-dependent Schrödinger equation for each block it is considered that the excitations are localized on the amide I modes, and the integration is performed in small time steps during which the Hamiltonian can be considered constant. In this way time-evolution matrices, $\mathbf{U}$, for each excitation manifold is obtained for the delays between the interactions at times denoted τ_0 to τ_4. This allows to calculate both linear response function governing the linear absorption [89].

$$S_{1D}(\tau_{10}) \quad = \quad -\left(\frac{i}{\hbar}\right) \langle \mu^{ge}(\tau_0) \mathbf{U}^{ee}(\tau_{01}) \mu^{eg}(\tau_1) \rangle. \tag{4.4}$$

Similarly the third-order response functions related to the Feynmann diagrams in Figure 3.8 are given by [83, 89, 108]:

$$\begin{aligned} S_{GB}^{(\vec{k}_I)}(t_3, t_2, t_1) = \quad & -\left(\tfrac{i}{\hbar}\right)^3 \langle \mu^{ge}(\tau_1) \mathbf{U}^{ee}(\tau_1, \tau_2) \mu^{eg}(\tau_2) \mu^{ge}(\tau_4) \\ & \mathbf{U}^{ee}(\tau_4, \tau_3) \mu^{eg}(\tau_3) \rangle_E \mathbf{\Gamma}(t_3, t_2, t_1) \\ S_{SE}^{(\vec{k}_I)}(t_3, t_2, t_1) = \quad & -\left(\tfrac{i}{\hbar}\right)^3 \langle \mu^{ge}(\tau_1) \mathbf{U}^{ee}(\tau_1, \tau_3) \mu^{eg}(\tau_3) \mu^{ge}(\tau_4) \\ & \mathbf{U}^{ee}(\tau_4, \tau_2) \mu^{eg}(\tau_2) \rangle_E \mathbf{\Gamma}(t_3, t_2, t_1) \end{aligned}$$

$$
\begin{aligned}
S_{EA}^{(\vec{k}_I)}(t_3,t_2,t_1) = & \left(\tfrac{i}{\hbar}\right)^3 \langle \mu^{ge}(\tau_1)\mathbf{U}^{ee}(\tau_1,\tau_4)\mu^{ef}(\tau_4) \\
& \mathbf{U}^{ff}(\tau_4,\tau_3)\mu^{fe}(\tau_3)\mathbf{U}^{ee}(\tau_3,\tau_2)\mu^{eg}(\tau_2)\rangle_E \Gamma(t_3,t_2,t_1) \\
S_{GB}^{(\vec{k}_{II})}(t_3,t_2,t_1) = & -\left(\tfrac{i}{\hbar}\right)^3 \langle \mu^{ge}(\tau_4)\mathbf{U}^{ee}(\tau_4,\tau_3)\mu^{eg}(\tau_3)\mu^{ge}(\tau_2) \\
& \mathbf{U}^{ee}(\tau_2,\tau_1)\mu^{eg}(\tau_1)\rangle_E \Gamma(t_3,t_2,t_1) \\
S_{SE}^{(\vec{k}_{II})}(t_3,t_2,t_1) = & -\left(\tfrac{i}{\hbar}\right)^3 \langle \mu^{ge}(\tau_2)\mathbf{U}^{ee}(\tau_2,\tau_3)\mu^{eg}(\tau_3)\mu^{ge}(\tau_4) \\
& \mathbf{U}^{ee}(\tau_4,\tau_1)\mu^{eg}(\tau_1)\rangle_E \Gamma(t_3,t_2,t_1) \\
S_{EA}^{(\vec{k}_{II})}(t_3,t_2,t_1) = & \left(\tfrac{i}{\hbar}\right)^3 \langle \mu^{ge}(\tau_2)\mathbf{U}^{ee}(\tau_2,\tau_4)\mu^{ef}(\tau_4) \\
& \mathbf{U}^{ff}(\tau_4,\tau_3)\mu^{fe}(\tau_3)\mathbf{U}^{ee}(\tau_3,\tau_1)\mu^{eg}(\tau_1)\rangle_E \Gamma(t_3,t_2,t_1)
\end{aligned}
\tag{4.5}
$$

where, $\Gamma(t_3,t_2,t_1)$ is the relaxation factor in which the vibrational lifetime T_1 is included and has the form:

$$
\Gamma(t_3,t_2,t_1) = e^{-\frac{t_3+2t_2+t_1}{2T_1}} \tag{4.6}
$$

The response functions Equation 4.4 and 4.5 are multiplied with exponential appodization functions corresponding to a vibrational lifetime of 1.8 ps [27,141]. The linear absorption is then obtained by a Fourier transform of Eq. 4.4 with respect to τ_{01}, while a two-dimensional Fourier transform with respect to the coherence times ($t_1 = \tau_{10}$ and $t_3 = \tau_{32}$) of Eq. 4.5 provides the 2D IR spectra [91], respectively. These coherence times were varied from 0 to 2.16 ps. The response functions were calculated from starting configurations along the trajectory separated by 2 ps, giving an ensemble average over a total of 500 realizations. The used coherence times result in a spectral resolution of 0.4 cm^{-1} as compared to the experimental resolution of 0.5 cm^{-1}. The spectral window in the calculations were set to the range 1400-1800 cm^{-1} allowing us to cover the entire experimental range from 1580-1715 cm^{-1}, which was truncated at these values to avoid contamination of the amide I region with contribution from amide II and other nearby vibrations.

4.2.4 Analysis

To provide the benchmark we need to determine the similarity between experimental and theoretical line shapes. We achieve this through the calculation of the spectral overlap [125, 141]:

$$S^{1D} = \sum_i (I(\omega_i) I_{ref}(\omega_i)) / \sqrt{\left(\sum_i I(\omega_i)^2\right) \times \left(\sum_i I_{ref}(\omega_i)^2\right)} \tag{4.7}$$

where $I(\omega_i)$ is the intensity of the theoretical spectra at a given frequency ω_i, for which the frequencies were shifted to maximize the overlap, while $I_{ref}(\omega_i)$ is the intensity of the experimental spectra for the same frequencies. A cubic spline interpolation was used, in order to determine the simulated spectral intensities at the same frequencies as experiment [142]. The shift of the theoretical spectrum, $\Delta\omega$, maximizing the spectral overlap was determined to distinguish between the ability to predict the spectral position and shape. For the linear spectra, which are always positive the values of the spectral overlap vary in a range between 0 to 1, where 1 indicates a perfect match between the theoretical and experimental line shapes, while a 0 corresponds to a complete mismatch between the spectral line shapes. As the linear spectra are always positive, negative values cannot be obtained.

For the 2D spectra a similar spectral overlap was determined using the spectral shift found for the linear spectra:

$$S^{2D} = \sum_{i,j} (I(\omega_i, \omega_j) I_{ref}(\omega_i, \omega_j)) / \sqrt{\left(\sum_{i,j} I(\omega_i, \omega_j)^2\right) \times \left(\sum_{i,j} I_{ref}(\omega_i, \omega_j)^2\right)} \tag{4.8}$$

For the 2D IR spectra the values of the spectral overlap vary between a range of -1 to 1. Where negative values may occur if absorption and bleach regions of the spectra have been interchanged.

4.3 RESULTS

4.3.1 The Linear Spectra

The experimental linear absorption spectra of the three proteins presented in Figure 4.2 are taken from Ref. 75 . The spectra are characterized by their peak positions and lineshape, which are related with secondary structure. Lysozyme has the narrowest spectra and the highest peak frequency, as a consequence of the high content of α-helices (52%) [77, 115]. Concanavalin A has the broadest spectrum and the lowest frequency peak position. A shoulder at approximately 1680 cm^{-1} is arising from the β-sheets presence. Ribonuclease A has a mixture of α-helix and β-sheet content.

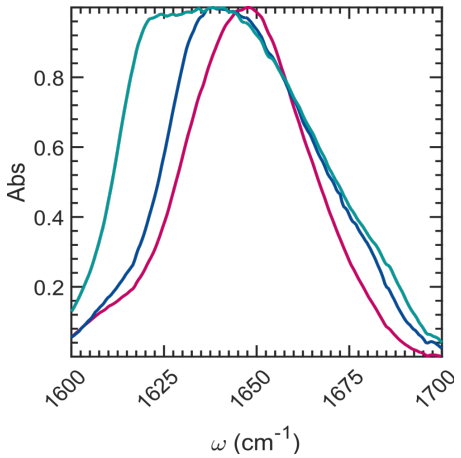

Figure 4.2: The experimental linear infrared spectra of Lysozyme (purple), Ribonuclease A (blue), and Concanavalin A (aquagreen). All the spectra were normalised with respect to maximum intensity.

As described in the methods section, in total nine models were used to extract the time dependent Hamiltonian Eq. (6.1) for both force fields (four for OPLS-AA and five for CHARMM-27). The linear spectra are shown in Figure 4.3. To quantify deviations in the peak positions, the frequency shift between experiment and theory was measured, by shifting the frequencies of theoretical spectra to maximize the spectral overlap Eq. (4.7). The numbers are given in Table 4.1, and a graphical representation is given in Figure 4.4.

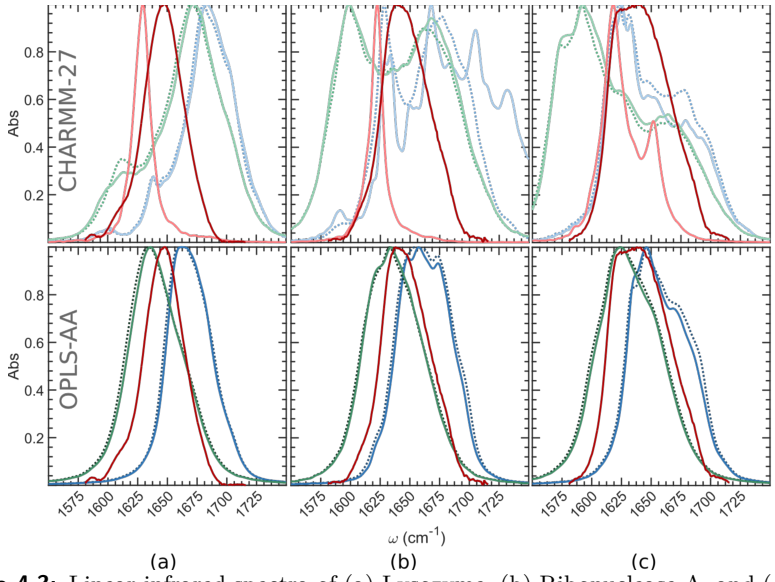

Figure 4.3: Linear infrared spectra of (a) Lysozyme, (b) Ribonuclease A, and (c) Concanavalin A, using the OPLS-AA and CHARMM-27 for MD simulation in combination with different electrostatic mappings and coupling models. The spectrum in dashed dark blue is for the Jansen TCC combination, and the light blue is the spectrum obtained with Jansen TDC combination. The spectra in dashed dark green and light green colors are the ones calculated with the Skinner TCC and Skinner TDC combination, respectively. The spectrum in pink is the one for Tokmakoff TCC combination. All experimental spectra are plotted in red color. All spectra have been normalised with respect to maximum intensity.

For the Jansen frequency map the peak positions are overestimated requiring a -17 cm^{-1} shift of the theory data to match experiment, in good agreement with previous findings for Trpzip2, where at -20 cm^{-1} shift was used [143]. For the Skinner frequency map mostly blue shifts are needed.

Model		Lys	RNseA	ConA	Average	s.d.
OPLS-AA	Jansen/TCC	-19.9	-15.3	-16.0	-17.1	2.0
	Jansen/TDC	-20.7	-15.4	-15.8	-17.3	2.4
	Skinner/TCC	9.5	11.4	10.1	10.3	0.8
	Skinner TDC	8.2	11.0	9.4	9.6	1.2
CHARMM-27	Jansen/TCC	-38.4	-24.8	0.6	-20.9	16.1
	Jansen/TDC	-39.7	-35.1	0.9	-24.6	18.1
	Skinner/TCC	-27.2	-24.7	46.7	-1.7	34.3
	Skinner/TDC	-26.1	-26.2	45.5	-2.3	33.8
	Tokmakoff/TCC	17.9	21.2	11.4	16.3	4.7
Amber99SB-ILDN	Jansen/TCC	-15.7	-15.6	-20.3	-17.2	2.2
	Skinner/TCC	15.2	13.8	11.1	13.4	1.7
	Tokmakoff/TCC	11.6	13.1	9.5	11.4	1.5
Gromos-54a7	Jansen/TCC	-17.3	-19.6	-26.1	-21.0	3.7
	Jansen/TDC	-18.8	-21.5	-26.8	-22.4	3.3
	Skinner/TCC	10.3	6.4	2.2	6.3	3.3
	Skinner/TDC	8.5	5.1	0.9	4.8	3.1
	Tokmakoff/TCC	19.3	15.3	6.7	13.8	5.3
	Tokmakoff/TDC	-9.9	-8.9	-11.3	-10.0	1.0

Table 4.1: The frequency shifts of the theoretical spectra applied to maximize the spectral overlap of the linear spectra. All numbers are given in cm^{-1}. The values for Amber99SB-ILDN and Gromos-54a7 are generated from the data of Ref. 125.

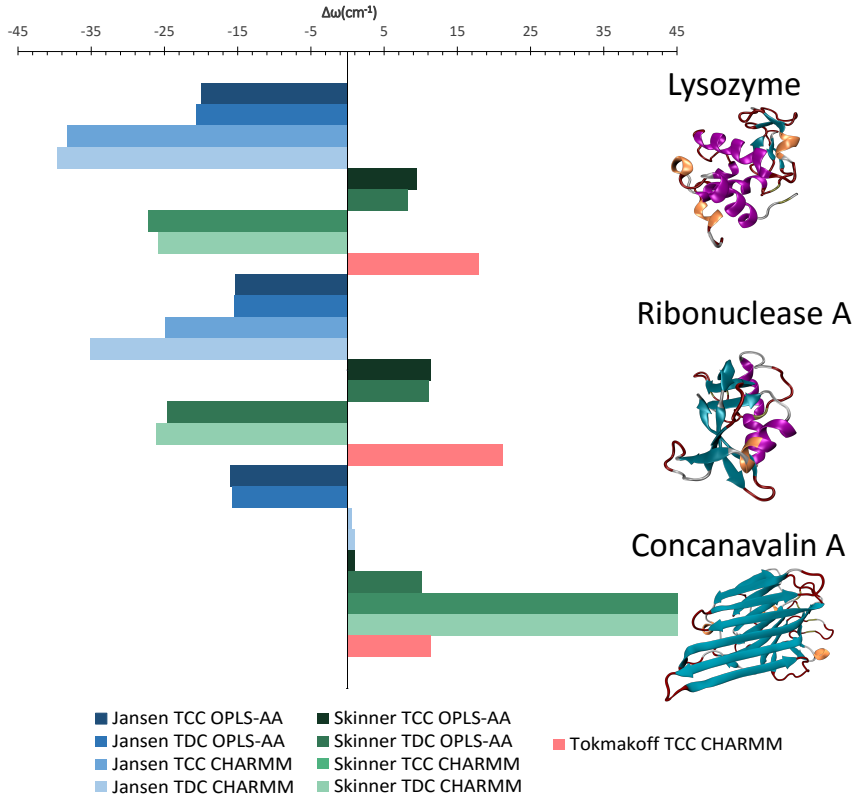

Figure 4.4: The frequency shifts ($\Delta\omega$) applied to maximize spectral overlap for the different simulation protocols.

For Lysozyme the linear spectra is dominated by a single peak, while for Ribonuclease A, and Concanavalin A, multiple peaks or shoulders are present. For the CHARMM-27 force field a general correlation between secondary structure and frequency shift was found, where the required shift become more positive as the percentage of β-sheet increases. The Tokmakoff map, which was developed for the CHARMM-27 force field exhibit the smallest deviations for this force field. These trends are force field dependent, and opposite to the findings for the Gromos-54a7 and Amber-99SB-ILDN force fields in Ref. 125, while for OPLS-AA the needed shift is generally independent of the different secondary structures, thus, allowing for a correction by a systematic shift. This effect is also recognized by the small standard deviations (s.d) of the frequency shifts reported in Table 4.1, which also summarize the data from Ref. 125.

Model		Lys	RNseA	ConA	Average	s.d.
OPLS-AA	Jansen/TCC	0.991	0.988	0.994	0.991	0.002
	Jansen/TDC	0.994	0.993	0.996	0.994	0.001
	Skinner/TCC	0.972	0.984	0.990	0.982	0.008
	Skinner/TDC	0.978	0.987	0.992	0.987	0.006
CHARMM-27	Jansen/TCC	0.981	0.869	0.603	0.818	0.158
	Jansen/TDC	0.982	0.824	0.949	0.91	0.068
	Skinner/TCC	0.944	0.824	0.937	0.901	0.055
	Skinner/TDC	0.953	0.831	0.935	0.906	0.054
	Tokmakoff/TCC	0.906	0.817	0.902	0.875	0.041

Table 4.2: The maximized spectral overlap (Eq. 4.7) between the theoretical FTIR and the experimental FTIR.

The spectral overlap for the linear spectra Eq. (4.7), given in Table 4.2, were calculated to evaluate the performance of the simulation protocols for predicting the lineshapes. The maximum spectral overlap is found for the Jansen/TDC/OPLS-AA combination associated with the minimum value of the standard deviation (s.d.). The average spectral overlap value is 0.99. The Tokmakoff/TCC/OPLS-AA combination has the lowest value for the coefficient overlap, while, the values for the Skinner combinations are all greater than 0.96. The quality of these values was evaluated by imposing two limits, Figure 4.5, where the black and the red lines represent the highest and lowest value for the spectral overlaps calculated between the experimental spectra of the different proteins. The lowest spectral overlap was obtained for the combination

Lysozyme and Concanvalin A, while the highest was obtained between Myoglobin and Ribonuclease A. Ideally the spectral overlap should, thus, be better than the highest of these values as the simulation protocols can then be expected to be useful to distinguish between the proteins with the most similar spectra. If the spectral overlap is lower than the lowest the simulation protocol cannot be expected to differentiate between the proteins with the most different spectra. The spectral overlap for the Jansen/TDC/OPLS-AA combination has values above the highest limit. The Tokmakoff/TCC/CHARMM-27 combination has values significantly below the lowest limit. The spectra with this protocol has a narrower distribution of the site frequencies, resulting in the presence of multiple peaks, and consequently the associated spectral overlap is smaller Figure 4.3. Our findings are in accordance to what has been reported by [125], in which the best combination was found for an all atom force field (Skinner/TCC/Amber-99SB-ILDN). However, our results obtained with OPLS-AA for the frequency shift show a lower s.d., than the ones reported in Ref. 125, meaning that the errors obtained with this force field are more systematic and possible to compensate by a fixed frequency shift. The spectral overlap, for OPLS-AA are also higher resulting in a better prediction of the lineshape by the use of this force field. The effect of the average shift on the spectral overlap was evaluated, and displayed in Table 4.3. This was achieved by shifting the spectra using the average shift and recalculating the spectral overlap for this frequency shift. These overlaps are generally just slightly lower than the ones obtained with the maximized shift. This allows choosing the best simulation protocol for future simulations, when a systematic shift according to Table 4.1 is performed to account for the error in the predicted peak position.

Model		Lys	RNseA	ConA	Average	s.d
OPLS-AA	Jansen/TCC	0.985	0.988	0.994	0.989	0.004
	Jansen/TDC	0.985	0.992	0.995	0.991	0.004
	Skinner/TCC	0.972	0.984	0.991	0.982	0.007
	Skinner/TDC	0.977	0.987	0.992	0.986	0.006
CHARMM-27	Jansen/TCC	0.799	0.867	0.820	0.829	0.030
	Jansen/TDC	0.845	0.807	0.762	0.805	0.030
	Skinner/TCC	0.726	0.766	0.721	0.738	0.020
	Skinner/TDC	0.711	0.757	0.725	0.731	0.019
	Tokmakoff/TCC	0.901	0.816	0.821	0.845	0.039
Amber99SB-ILDN	Jansen/TCC	0.985	0.991	0.989	0.988	0.002
	Skinner/TCC	0.975	0.989	0.989	0.984	0.007
	Tokmakoff/TCC	0.947	0.943	0.973	0.954	0.013
Gromos-54a7	Jansen/TCC	0.975	0.982	0.980	0.979	0.003
	Jansen/TDC	0.977	0.989	0.979	0.982	0.005
	Skinner/TCC	0.973	0.991	0.983	0.982	0.007
	Skinner/TDC	0.980	0.994	0.982	0.985	0.006
	Tokmakoff/TCC	0.930	0.918	0.903	0.917	0.011
	Tokmakoff/TDC	0.991	0.996	0.988	0.992	0.003

Table 4.3: The spectral overlap between the theoretical FTIR and the experimental FTIR calculated with the average frequency shifts. The values for Amber99SB-ILDN and Gromos-54a7 are generated from the data of Ref. 125.

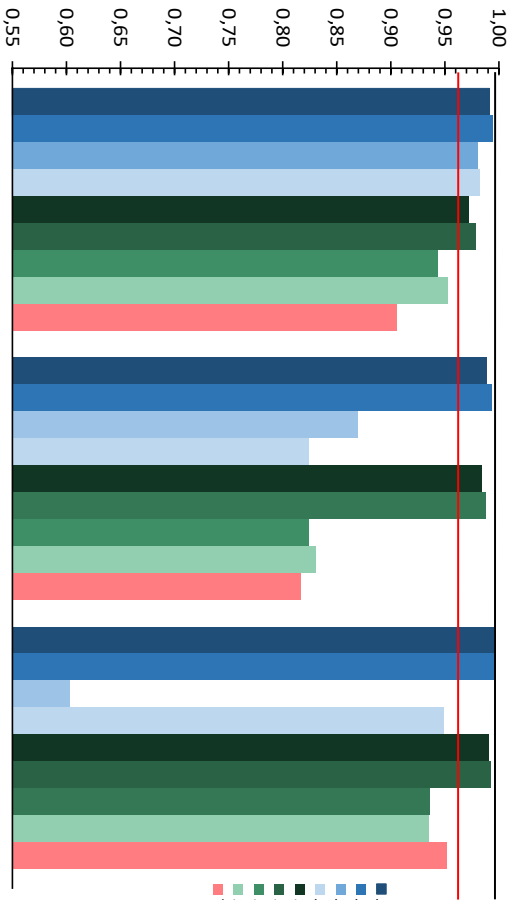

Figure 4.5: The spectral overlaps between the experimental and theoretical linear spectra. The upper limit (black solid) is given by the overlap of the experimental spectra of Myoglobin and Ribonuclease A, while the lower limit (red solid) is given the overlap between experimental spectra of Lysozyme and Concanavalin A.

4.3.2 The 2D IR Spectra

2D IR are expected to be more sensitive to structure and dynamics, and, therefore, are a more sensitive benchmark and generally more difficult to model. The simulated and experimental 2D IR spectra are given in Figure 4.6. This include the OPLS-AA force field combined with the Jansen and Skinner maps, which gave good predictions of the linear absorption for this force field. The CHARMM-27 force field is included as well in the combination with the Tokmakoff map, which was the only one giving acceptable linear absorption spectra for this force field. Clear differences are observed between 2D IR spectra predicted with the different simulation protocols. The spectral overlap was calculated using Eq. (4.8) after shifting the spectra to maximize the spectral overlap for the linear absorption. We, thus, quantify the 2D IR lineshapes and not the ability to predict correct peak positions. A bilinear interpolation was used, to obtain the intensities of the simulated spectra at the same frequencies as the experimental ones. The resulting spectral overlaps are shown in Table 4.4, and Figure 4.7. Here, the red and black line, again, represent the lowest and highest limit for the spectral overlap, respectively, and were calculated using the experimental spectra with most similar (Myoglobin and Lysozyme), and different line shape (Lysozyme and Concavalin A). The highest average spectral overlaps were found for the Skinner mapping combinations, with an average coefficient overlap of 0.89. Furthermore, the associated s.d. is low, meaning that the quality of the predictions are consistently high. The 2D IR spectra simulated with the Tokmakoff and Jansen mapping combinations have the presence of multiple peaks, which are due to a too narrow distribution of the site frequencies. The Skinner map gives rise to broader spectra, in which the peaks are more elongated along the diagonal, providing better agreement with experiment. Furthermore, the spectral overlaps for all Skinner mapping combinations are above the lowest limit, meaning that both combinations are suitable for simulating 2D IR spectra using OPLS-AA.

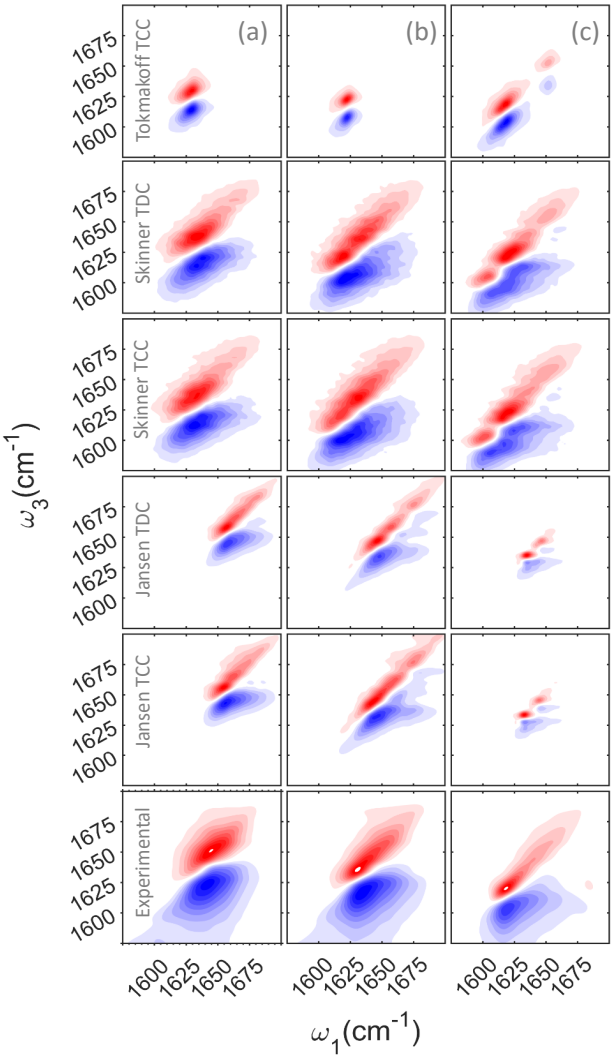

Figure 4.6: Two dimensional infrared spectra of the proteins: (a) Lysozyme, (b) Ribonuclease A, (c) Concanavalin A, simulated with the OPLS-AA force field with the exception of the spectra with the Tokmakoff map which were calculated using the CHARMM-27 force field. The contour lines are equidistant, and separated by 10% of the maximum intensity, and all the spectra have been normalised with respect to the maximum intensity.

Model		Lys	RNseA	ConA	Average	s.d.
OPLS-AA	Jansen/TCC	0.797	0.819	0.663	0.760	0.069
	Jansen/TDC	0.767	0.792	0.654	0.738	0.060
	Skinner/TCC	0.921	0.939	0.841	0.887	0.046
	Skinner/TDC	0.920	0.929	0.833	0.894	0.041
CHARMM-27	Tokmakoff/TCC	0.675	0.551	0.695	0.641	0.063

Table 4.4: The spectral overlap (Eq. 4.8) between the theoretical 2D IR and the experimental 2D IR.

The 2D IR spectral overlaps are lower for the Jansen mapping combinations than the ones obtained in Ref. 125 for combination of the Jansen map with the Amber force field. However, the Skinner combinations have a higher spectral overlap than what has been reported there, meaning that the prediction of the 2D IR lineshapes using OPLS-AA with these combination are the best so far.

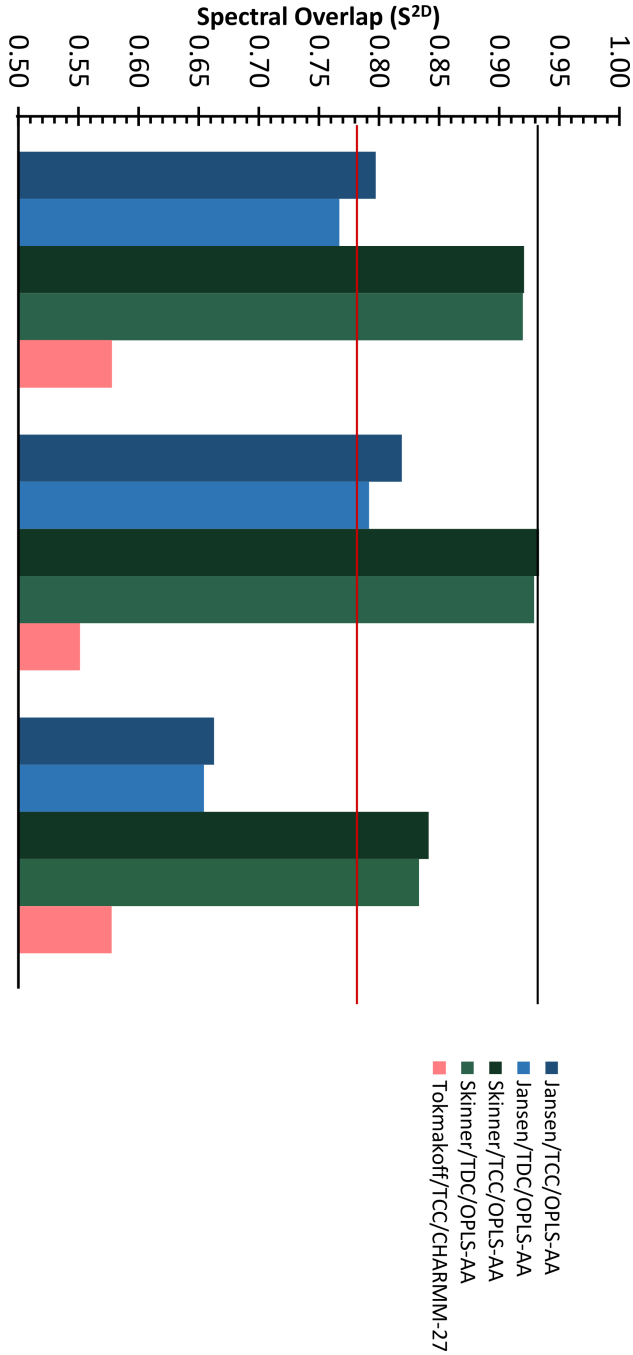

Figure 4.7: The spectral overlaps between the experimental and theoretical 2D IR spectra. The upper limit (black solid) is given by the overlap of the experimental spectra of Myoglobin and Ribonuclease A, while the lower limit (red solid) is given the overlap between experimental spectra of Lysozyme and Concanavalin A.

Model		Lys	RNseA	ConA	Average	s.d.
OPLS-AA	Jansen/TCC	0.797	0.807	0.665	0.756	0.065
	Jansen/TDC	0.747	0.761	0.674	0.727	0.038
	Skinner/TCC	0.923	0.932	0.841	0.898	0.041
	Skinner/TDC	0.919	0.955	0.832	0.902	0.052
CHARMM-27	Tokmakoff/TCC	0.656	0.573	0.484	0.571	0.070
Amber99SB-ILDN	Jansen/TCC	0.802	0.842	0.710	0.785	0.055
	Skinner/TCC	0.865	0.898	0.774	0.843	0.051
	Tokmakoff/TCC	0.771	0.763	0.749	0.761	0.009
Gromos-54a7	Jansen/TCC	0.787	0.828	0.538	0.718	0.128
	Jansen/TDC	0.781	0.815	0.515	0.704	0.134
	Skinner/TCC	0.833	0.901	0.667	0.800	0.098
	Skinner/TDC	0.833	0.883	0.647	0.788	0.102
	Tokmakoff/TCC	0.689	0.749	0.456	0.631	0.126
	Tokmakoff/TDC	0.850	0.860	0.508	0.739	0.164

Table 4.5: The spectral overlap (Eq. 4.8) between the theoretical 2D IR and the experimental 2D IR. The values for Amber99SB-ILDN and Gromos-54a7 are generated from the data of Ref. 125.

The overlaps evaluated using the average frequency shift for the 2D IR spectra are presented in Table 4.5. These overlaps are similar to the ones obtained with the maximized shift. allows choosing the best simulation protocol for future simulations, when a shift maximizing the overlap cannot be performed and a systematic shift according to Table 4.1 is applied instead.

4.3.3 Discussion

In this work, two force fields, CHARMM-27 and OPLS-AA, in combination with three electrostatic maps, and two coupling models, were tested for lineshape and peak position prediction of the amide I band. The force fields were not developed for spectral simulations, which may be a source of discrepancies between the simulated and experimental spectra. In previous work, [125], the Skinner/TCC/Amber-99SB-ILDN combination was reported to give the best results for this type of simulation protocol, with average overlap of 0.986 for FTIR and 0.862 for 2D IR. Our results are in agreement with what was obtained before, in which both Jansen and Tokmakoff electrostatic maps seem to underestimate the spectral broadening, giving rise to narrow spectra with

multiple peaks. In both cases, the Skinner map shows the best results, giving rise to a more accurate prediction of the spectral line shape.

It is important to observe that the Jansen and Skinner maps do not work properly for the CHARMM-27 force field. This is most likely due to the differences in the electrostatics predicted by the force fields may be further affected by the water model combination. This clearly demonstrates that validation, as we present in this Chapter, is crucial before the application of new simulation protocols to predict spectra. We observe that for the CHARMM-27 force field the Tokmakoff map is the only one giving reasonable predictions of the peak positions, while the spectral widths are underestimated. This can probably be explained by the fact that the Tokmakoff mapping [30] was fitted to match the peak positions in dipeptides, while the peak widths were not explicitly included in the fitting. It may, thus, be possible to construct a better map for CHARMM-27 by including these in the fitting procedure.

Our results show that many models used for spectral simulation should be possible to improve, either by developing more accurate models, or by adapting the existing models to the most popular force fields. In the present work a fixed anharmonicty of 16 cm^{-1}, and a lifetime of 1.8 ps is used for all spectral simulations. These parameters could potentially be improved, however, considering the computational cost of the present benchmarking procedure we limited ourselves to the values deduced from previous studies [23,27]. Furthermore, the time dependent Hamiltonian is constructed from short MD simulations, which may neglect protein flexibility at much longer timescales. This can be overcome by testing longer trajectories in order to sample the protein conformational space better. This will potentially improve the protocols that currently predict too narrow spectra, once sampling over a more inhomogeneous distribution of structures can be expected to lead to broader spectra. Here, the SPC/E water [132] is used to solvate protein structures as was done in a previous benchmarking paper [125]. Commonly the TIP3P [144] is used with the OPLS-AA force field [130]. The choice of SPC/E was made as this model predicts the dynamical properties of water as diffusion coefficients, which are crucial for the 2D IR, better than most other water models [145, 146]. One could consider to improve the simulation protocol, by testing other water models [144,145,147,148], including recently polarizable water models [149–153], or protein force fields [154–161]. The polarizable models are, however, generally computationally demanding and not implemented in all MD packages.

In the present study we ignored the amide vibrations known to be present in the glutamine and asparagine side chains. So far only one mapping exists

for these side chain vibrations [29]. These vibrations generally absorb at higher frequencies than the backbone vibrations and can be expected to contribute to the inhomogeneity of the spectra. It will, thus, be desirable to include these vibrations in future studies.

The protocols give the worst results for Concanavalin A. We do not know the fundamental reason behind this. In principle it may be that the nearest neighbour coupling model [120] does not describe β-sheets that well. It may be due to the neglect of side chain vibrations. Alternatively, the experimental spectra may depend sensitively on ion and protein concentrations and pH in a way not well accounted for by the MD simulations. The latter is substantiated by the fact that spectra of Concanavalin A at higher protein concentrations [79] are somewhat different than for the new experimental spectra used here [75]. This was also discussed in more detail in a previous paper [125].

4.4 CONCLUSIONS

We benchmarked spectral simulation protocols for amide I spectroscopy of proteins using the CHARMM-27 and OPLS-AA force fields and three proteins with different secondary structure content. The results were compared with the findings of a recent benchmark study for Gromos-54a7 and Amber-99SB-ILDN. The quality of the simulation protocols were in both cases quantified using the frequency shift needed to best match the linear absorption spectra with experiment, and spectral overlaps for linear and 2D IR spectra. The current benchmarking should help users to choose the optimal simulation protocol, thus, increasing the quality of the simulated spectra and improving interpretations of the amide I band of proteins based on simulations. Eventually, this will pave the way for using 2D IR spectroscopy for benchmarking force fields like has been done with NMR [90]. We do find that the current models for spectral simulation leave points for further improvement. In particular, we think that in the near future one should combine a polarizable force field for proteins, and water, to evaluate the possibility of improving the predictions of the simulated spectra with such models.

The best model was shown to be the Skinner frequency map combined with the transition dipole coupling model and the OPLS-AA force field. This combination exhibited a very systematic error for the average frequency shift requiring a $+10$ cm^{-1} frequency shift to match experiment. For the linear absorption the

best models performed very similarly, while the Skinner/TDC/OPLS-AA combination performed clearly better for the 2D IR spectra. The Jansen frequency map performed well for predicting spectral positions and linear absorption, but gave too sharp peaks in the 2D IR spectra. We, thus, conclude that these combinations are currently the best to use for simulating the amide I band of proteins.